Hyperspherical Learning in Multi-Label Classification

Bo Ke[1], Yunquan Zhu[1], Mengtian Li[2], Xiujun Shu[1], Ruizhi Qiao[1], and Bo Ren[1]

[1] Tencent YouTu Lab
{boke, yunquanzhu, ruizhiqiao, timren}@tencent.com, shuxj@mail.ioa.ac.cn
[2] East China Normal University
mtli@stu.ecnu.edu.cn

Abstract. Learning from online data with noisy web labels is gaining more attention due to the increasing cost of fully annotated datasets in large-scale multi-label classification tasks. Partial (positive) annotated data, as a particular case of data with noisy labels, are economically accessible. And they serve as benchmarks to evaluate the learning capacity of state-of-the-art methods in real scenarios, though they contain a large number of samples with false negative labels. Existing (partial) multi-label methods are usually studied in the Euclidean space, where the relationship between the label embeddings and image features is not symmetrical and thus can be challenging to learn. To alleviate this problem, we propose reformulating the task into a hyperspherical space, where an angular margin can be incorporated into a hyperspherical multi-label loss function. This margin allows us to effectively balance the impact of false negative and true positive labels. We further design a mechanism to tune the angular margin and scale adaptively. We investigate the effectiveness of our method under three multi-label scenarios (single positive labels, partial positive labels and full labels) on four datasets (VOC12, COCO, CUB-200 and NUS-WIDE). In the single and partial positive labels scenarios, our method achieves state-of-the-art performance. The robustness of our method is verified by comparing the performances at different proportions of partial positive labels in the datasets. Our method also obtains more than 1% improvement over the BCE loss even on the fully annotated scenario. Analysis shows that the learned label embeddings potentially correspond to actual label correlation, since in hyperspherical space label embeddings and image features are symmetrical and interchangeable. This further indicates the geometric interpretability of our method. Code is available at https://github.com/TencentYoutuResearch/MultiLabel-HML.

Keywords: Multi-Label Classification, Partial Labels, Label Correlation

1 Introduction

Multi-label classification has a wide range of applications in generic scenarios, including medical image processing [2], pedestrian attribute recognition [56] and

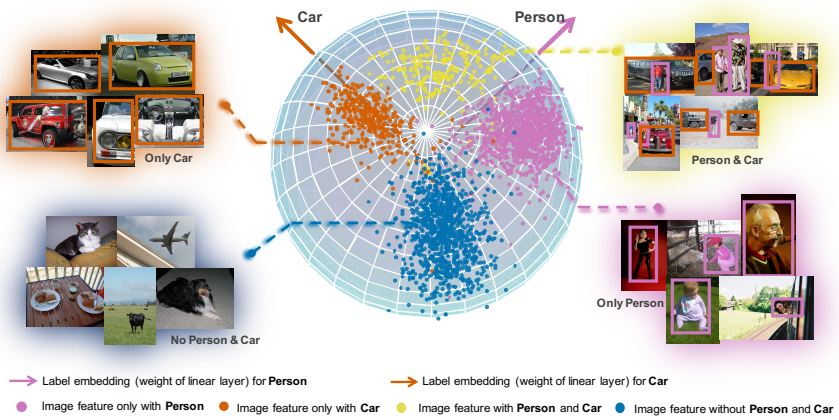

Fig. 1. The Overview of learning multi-label with *Person* and *Car* in 3D hyperspherical space. In this case, the cosine similarity between multiple label embeddings and image features is used as the metric for classification. With optimization, images with the same labels are clustered together in a hypersphere. Best viewed in color.

image retrieval [53]. However, it is not easy to construct a large-scale multi-label dataset by manual annotation. Human annotators need to learn about a large number of nameable labels and assign positive labels accurately. Due to the limitation of human knowledge and fatigue, human annotators tend to skip some positive labels, causing false negatives to occur. In order to alleviate human labour, another effective strategy is to generate datasets from noisy web labels [7]. In this way, the dataset is built by crawling web images by using the labels as queries [26]. Therefore, only one correct label can be obtained for each image; other unknown labels, whether they are actually present or not, are considered as negatives [12]. Those false negatives in annotations inevitably lead to the degradation in the generalization capability of the network. To address the problem of missing positive labels, our work focuses on studying false negative labels under three multi-label scenarios, consisting of single positive labels, partial positive labels and full labels.

In full labels scenarios, the method on asymmetric loss [38] obtains gains on performance by increasing the weights of hard samples. For noisy data, existing approaches are more concerned with correcting inaccurate labels by pseudo-labeling [36][45][1]. However, there is no unified perspective that can comprehensively handle different proportions of mislabeled samples in the multi-label task. To further explore this issue, we propose to model the multi-label task in the hyperspherical space, as shown in Fig. 1.

The motivation of this hyperspherical reformulation lies in two aspects. First, the relationship between label embedding and image feature is expressed in terms of the angle on the sphere, independent of the magnitude of label embedding or image feature. In multi-label classification, the inter-class distribution is highly imbalanced, which easily leads to poor generalization. With the metric of nor-

malized embeddings and features, the contributions of each class are treated as equal, thus avoiding overfitting [43]. Second, compared to the Euclidean space, the similarity in hyperspherical space is bounded. It is convenient to use metric learning to equalize the weights of samples in bounded space [9]. Especially in noisy data, the weight of the false negatives is amplified during training. These problems can be naturally addressed by incorporating an angular margin based loss function in hyperspherical space. Comprehensively, our work proposes to learn multi-label in hyperspherical space, which is proved to be valid for multi-label datasets with different levels of noisy samples.

It is well known that significant improvement in multi-label classification can be achieved by exploiting label correlations [21]. With the introduction of hyperspherical space, an additional benefit is that the model implicitly learns label correlation. The cosine similarity between label embedding and image feature is used to determine if the label is attached to the image. Since label embeddings and image features are symmetrical and interchangeable in hyperspherical space, the cosine similarity between label embeddings inherently illustrates the label correlation.

In summary, the main contributions in this paper can be concluded as

- For the first time, the multi-label task is modelled in hyperspherical space, which provides a new perspective for the field of multi-label learning.
- A novel angular margin based multi-label loss is presented to handle false negatives in multi-label classification which cannot be dealt with in Euclidean space due to unbounded distance.
- Our method is evaluated under the scenarios of single positive labels, partial positive labels and full labels, demonstrating the effectiveness of the proposed method.
- The geometric interpretability of the method is further explained by an analysis of label correlations.

2 Related Works

2.1 Learning from Noisy Labels

In multi-label tasks, incompletely labelled data will inevitably be used for training. To analyze the multi-label performance with noisy labels, previous works propose a variety of scenarios [20][35][17]. One of the most commonly used is *weak label*, which assumes the known labels are proper labels while unknown labels are regarded as negatives [39][10][49]. In order to reduce the impact of false negatives in *weak label*, unknown labels are not used for training in the *partial labels* scenario [12][8][6][19]. In *positive-unlabeled learning*, only some positive and unlabeled samples are accessed [27][13]. More strictly, *Single positive labels* proposes the setup that only one single positive label is available for each image at training time [7][51]. Our work mainly investigates learning multi-label models with single and partial positive labels, which is the most practical and economically accessible scenario of noise web labels, as web images are more

likely to be published with incomplete correct labels than with completely correct labels or with incorrect labels. In the *single positive labels* scenario, the performance gain of existing methods [7] is only seen on limited datasets, while our proposed method consistently achieves the state-of-the-art performance on multiple datasets, including VOC12 [14], COCO [28], CUB-200 [41] and NUS-WIDE [5]. Our method also obtains significant improvement in the scenarios of partial positive labels and full labels. Moreover, we compare with more noisy single-label classification methods in supplementary material.

2.2 Hyperspherical Learning

Hyperspherical learning has made great progress in face recognition in recent years [30][47]. NormFace [43] first introduces training embedding using normalized features in face verification. Sphereface [29] proposes to learn embeddings using large angular margin in open-set face recognition. CosFace [44], Additive margin softmax [42] and ArcFace [9] further improves the form of angular margin to stabilize the training. Learning features in hyperspherical space is also popular in person re-identification. A deep cosine similarity metric is firstly used to achieve better generalization on the test set in person re-identification [48]. Another simple but strong baseline with normalized softmax is also proposed to reduce the difficulty of optimization in person re-identification [32][33][15]. All these works aim to learn representations on hyperspherical space and prove that angular metric is crucial to the generalization in retrieval. Inspired by these observations, we propose to learn multi-label classification in hyperspherical space. Our method differs from existing methods in two aspects. First, while existing methods focus on learning features for retrieval tasks like face recognition and person re-identification, our approach focuses on classification. Second, previous approaches on hyperspherical learning are limited to single class tasks using normalized softmax, while multi-label task, to the best of our knowledge, remains unexplored in hyperspherical learning.

2.3 Label Correlation

It is well known that exploiting label correlations is crucial for multi-label classification [21]. Some existing approaches assume that label correlation is shared only in a local group of instances [18], while others deal with missing labels by exploiting both global and local label correlations [57][52]. Graph convolution network (GCN) is naturally suitable to build label correlation, and thus has gained much attention in multi-label tasks [46][25][34]. ML-GCN [4] utilizes graphs to propagate prior label representations, such as word embeddings, in learning classifiers. ADD-GCN [50] learns content-aware category representations without using an external word embedding for graph construction. Recently, vision transformer has gained popularity in the field of image recognition [11][40][31][37][3]. Transformers are also used to construct complementary relationships in multi-label classification by exploring structural relation graph and semantic relation graph [54]. M3TR [55] presents a linguistic guided enhancement method to enhance

the high-level semantics. These works demonstrate the potential of transformers for building label correlation [24]. Our approach provides an alternative way to implicitly learn label correlation in multi-label classification by applying constraints in hyperspherical space, where the label embedding should maximize the similarity with corresponding image features and minimize the similarity with other labels if existing label conflicts.

3 Method

In this section, we thoroughly introduce multi-label classification into hyperspherical feature space. In Sec. 3.1, the preliminary study demonstrates the multi-label classification in Euclidean space. In Sec. 3.2, normalized sigmoid function is proposed to learn multi-label in hyperspherical space. In Sec. 3.3, a modified variant, named margin based sigmoid function, is proposed to handle false negatives in noisy data. In Sec. 3.4, we enable the hyperparameters to be adaptive to a wide variety of datasets. In Sec. 3.5, label correlation is illustrated in hyperspherical space.

3.1 Preliminaries

In multi-label classification, each image needs to be determined whether it belongs to the labels in given sets. Unlike single-label image classification in ImageNet [23], the number of output labels in the multi-label setting may be one, many, or none. The most common practice is to use multiple binary classifiers. Each classifier determines whether the corresponding label exists in the image. Assume that we need to optimize a multi-label classifier with N labels. The multi-label problem could be optimized with binary cross-entropy loss with sigmoid activation $\sigma(z) = 1/(1 + e^{-z})$.

$$\mathcal{L}_{\mathrm{BCE}}(p, \mathcal{Y}) = -(\sum_{i \in \mathcal{Y}}^{N} \log(p_i) + \sum_{i \notin \mathcal{Y}}^{N} \log(1 - p_i)) \tag{1}$$

where $\mathcal{Y}$ is the set of the true labels in corresponding image, p are the probabilities for the N labels. In Euclidean space, the probabilities are expressed as p^e,

$$p^e(W, b, x) = \sigma(Wx + b) \tag{2}$$

where W are learnable weights of N binary classifiers with the shape of *(N, L)*, b are learnable biases with the shape of *(N,)* and x is the L-dimension feature of the image. In Eq. 2, logits are expressed as the inner product of weights and features in Euclidean space. The sigmoid function then maps the logits into per-class probabilities within the range of $[0, 1]$. With larger probability p_i^e, the sample is more likely to belong to the i class, and vice versa. As $p^e = \sigma(\|W\| \, \|x\| \cos \angle(W, x) + b)$, the logits are not only related to the angle between classifier weights and features, but also affected by the weights norm and bias of

classifiers. We have concerns about learning in Euclidean space in two aspects. First, we suspect that over-optimization of weights norm leads to imbalanced learning between labels. Second, the angle is the better measure of similarity between weights and features compared to the norm. The objective function focusing on the angle is proposed in Sec. 3.2.

3.2 Learning in Hyperspherical Space

First, We utilize cosine similarity between the feature and classifier weights to minimize the binary cross-entropy loss. To simplify Eq. 2, we remove the modules that are not cross-correlated between W and x, such as bias b, weight norm $\|W\|$ and feature norm $\|x\|$, and thus propose normalized sigmoid function,

$$p^n(W, x) = \sigma(s * \frac{Wx}{\|W\| \, \|x\|}) \tag{3}$$

In Eq. 3, s is the scale factor of the cosine similarity. Different from the norms implicitly used in Eq. 2, the scale s is irrelevant to features x and weights W. It is used to rescale the cosine similarity to reach the saturation zone of sigmoid activation. We can treat the weights W as a collection of *label embeddings* in the perspective of the hypersphere. The similarity value ranges from -1 to 1 indicating the confidence of regarding the corresponding label in the given image. The overall loss function minimizes the angle between the feature and the corresponding positive label embedding and maximizes that between the feature and its negative label embedding.

There are two other advantages to learning multi-label in hyperspherical space. First, from the perspective of the hypersphere, we can efficiently conduct metric learning to handle noisy samples in multi-label classification, which is introduced in Sec. 3.3. Second, the correlation value between different label embeddings could be used to estimate the label correlation, which is illustrated in Sec. 3.5.

3.3 Learning from Single Positive Labels

To study the impact of noisy samples in multi-label classification, we analyze its simplest form, that is, the single positive labels scenario. In this problem, only one single positive label is known in each image; thus, unknown labels may be positive or negative in fact. A straightforward way is regarding all known labels as positive and all unknown labels as negative during training. We conventionally refer to samples with positive labels as *postives* and samples with negative labels as *negatives*. In this setting, the negatives are composed of true negatives and false negatives according to the ground truth.

We start with an experiment of full labels on VOC12, as shown in the left column of Fig. 2. In this experiment, no noisy labels are added to the training data; thus, all negatives are true negatives. Fig. 2 (a) illustrates the cosine similarity distribution of positives and negatives. The result shows less overlap

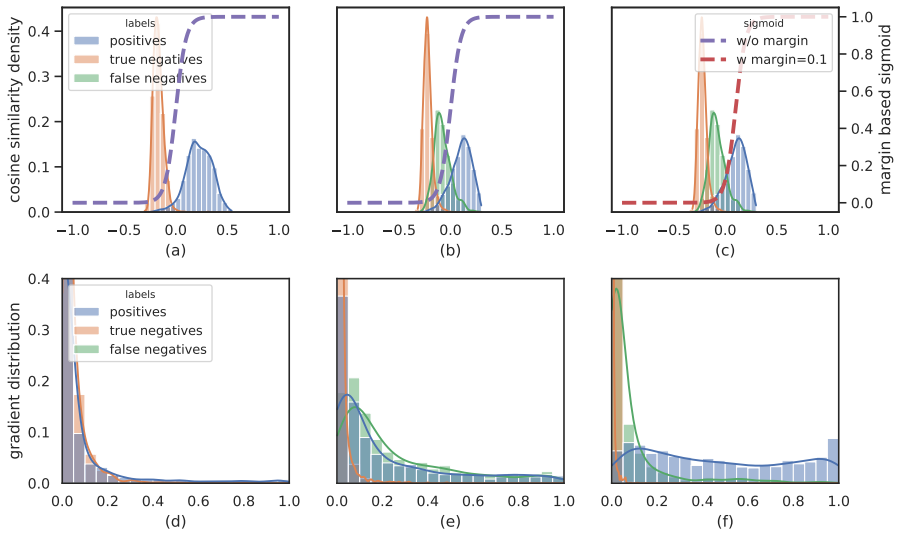

Fig. 2. The cosine similarity density (top rows) and the corresponding gradient distribution (bottom rows). The scenarios include normalized sigmoid in full labels (left column), normalized sigmoid in single positive labels (middle column), margin based sigmoid in single positive labels (right column). For the cosine similarity density, the color bar in the Y-axis represents the density in the corresponding similarity (X-axis). For the gradient distribution, the gradient magnitude is attached to the X-axis, while the color bar in the Y-axis represents the density. The actual labels of samples are distinguished by color.

between the distribution of positives and negatives and suggests that positives and negatives are easier to distinguish in the absence of noisy samples.

The middle column of Fig. 2 presents the experiment in single positive labels. The similarity distribution is close to that in full labels except for false negatives, which are partially overlapped with positives and true negatives in distribution, as shown in Fig. 2 (b). There is a conflict for false negatives. On the one hand, these mislabeled samples have visually similar patterns or textures to the positives. However, on the other, these samples are assigned to be negative, making them harder to distinguish. The gradient analysis on single positive labels is illustrated in Fig. 2 (e). Compared to the case of full labels in Fig. 2 (d), the gradients of false negatives still remain at high intensity when the training has converged. These incorrect gradients can mislead the network to converge to a non-optimal solution. We conclude that these hard false negatives cause the generalization gap between the fully-labeled dataset and the dataset with single positive labels.

The impact of those false negatives should be suppressed while keeping the contribution of the positives. To this end, we introduce angular margin on cosine similarity in our multi-label formulation, also named margin based sigmoid

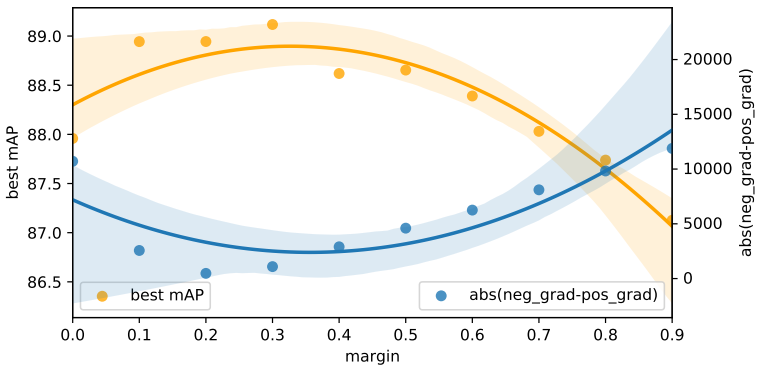

Fig. 3. The experiments with fixed margins in the VOC12 dataset. Orange line illustrates the best mAP in different margin. Blue line illustrates the absolute value of the accumulated gradient difference between the positives and negatives.

function,

$$p^m(W, x) = \sigma(s * \cos(\arccos(\frac{Wx}{\|W\| \, \|x\|}) + m)) \tag{4}$$

In Eq. 4, the margin m is added to the angle between classifier weights and features. The dotted lines in Fig. 2 show the variants of margin based sigmoid functions that map the cosine similarity to the probability. We further increase the margin from 0 to 0.1, as shown in the third column in Fig. 2. Observe that the activation function has been shifted to the right by 0.1. From the gradient analysis in Fig. 2 (f), the gradient of positives is enhanced while the gradient of negatives is weakened comparing to Fig. 2 (e). On the one hand, the harmful gradients from the false negatives are partially inhibited. On the other hand, the positives are activated to maximize the cosine similarity between class weights and their features. In multi-label classification, the classifier is the crucial part of both training and testing. To keep the consistency, we use the same activation function as Eq. 4 during training and testing.

3.4 Adaptive Learning

As discussed in Sec. 3.3, margin m optimizes the training by adjusting the ratio of positive and negative gradients. The smaller margin would produce a large number of gradients for negatives. Thus the absolute value of the accumulated gradient difference between the positives and negatives is significant as shown in Fig. 3. This imbalance of gradients results in poor performance. A similar degradation also happens when training with a large margin, where the accumulated gradients for positives are more extensive than those for negatives. It is vital to choose the appropriate margin to make the gradients of positives and negatives balanced. From Fig. 3, it can be seen that the appropriate margin in the VOC12 dataset is approximately between 0.2 and 0.4. However, it is tedious

to select the scale and margin through a large number of experiments with other datasets. Based on these observations, we further propose a gradient balanced loss function to learn adaptive scale and margin during training.

$$\mathcal{L}_{overall}(p, \mathcal{Y}) = \mathcal{L}_{\text{BCE}}(p, \mathcal{Y}) + \mathcal{L}_{adpt}(p, \mathcal{Y}) \tag{5}$$

A constraint on balancing gradients $\mathcal{L}_{adpt}$ works with binary cross-entropy loss in Eq. 5.

$$\mathcal{L}_{adpt}(p, \mathcal{Y}) = \left\| \sum_{i \in \mathcal{Y}}^{N} (1 - p_i) - \sum_{i \notin \mathcal{Y}}^{N} p_i \right\| \tag{6}$$

We aim to minimize the difference of the accumulated gradients between positives and negatives in a mini-batch, as in Eq. 6. Where $\sum_{i \in \mathcal{Y}}^{N}(1 - p_i)$ is the accumulated gradients for positives and the $\sum_{i \notin \mathcal{Y}}^{N} p_i$ is the accumulated gradients for negatives. Note that the probability p^a used in the adaptive loss is slightly different from p^m in BCE loss. In the adaptive loss, the gradients to weights and features are blocked, while the scale s^* and margin m^* are learnable parameters, as shown in Eq. 7.

$$p^a(W, x) = \sigma(s^* * \cos(\arccos(\mathbb{B}(\frac{Wx}{\|W\| \, \|x\|})) + m^*)) \tag{7}$$

Where $\mathbb{B}$ is the gradient blocking function. The adaptive loss focuses on balancing the gradient by learning the appropriate scale s^* and margin m^* without affecting the model weights W. The benefits of adaptive learning are in two aspects. First, adaptive learning removes manual hyperparameters searching on new datasets, such as COCO and NUS. Second, gradient equilibrium provides a better prior for model optimization and thus can boost the performance by adaptive learning.

3.5 Label Correlation

As in Fig. 1, the image features with label *person* is around the label embedding of *person* in 3D spherical space. The label embedding could be seen as the cluster center for samples with the same labels. Since label embeddings and image features are symmetrical and interchangeable in hyperspherical space, we could also replace the image features with label embeddings to compute correlation. Similar to Eq. 4, the relation between different label embeddings is formulated as Eq. 8, which could be used to estimate the label correlation.

$$Corr(i, j) = \sigma(s * \cos(\arccos(\frac{w_i w_j}{\|w_i\| \, \|w_j\|}) + m)) \tag{8}$$

Where w_i and w_j are the label embeddings of the ith label and jth label, respectively. The higher the value, the higher the correlation.

Table 1. Statistics of datasets.

Datasets	#Class	#Pos/Img	#Train Imgs	#Test Imgs
VOC12	20	1.44	5,717	5,823
COCO	80	2.92	82,081	40,137
CUB	312	31.47	5,994	5,794
NUS	81	1.89	150,000	60,260

Table 2. Experiments on single positive labels in mAP metric.

Method	VOC12	COCO	CUB	NUS
AN [7]	85.1	64.1	19.1	42.0
LS [7]	86.7	<u>66.9</u>	17.9	44.9
WAN [7]	86.5	64.8	<u>20.3</u>	<u>46.3</u>
EPR [7]	85.5	63.3	20.0	46.0
ROLE [7]	<u>87.9</u>	66.3	15.0	43.1
HML (Ours)	**89.1**	**70.7**	**21.1**	**46.7**

Table 3. Experiments on full labels in mAP metric.

Method	VOC12	COCO	CUB	NUS
BCE [7]	89.1	75.8	32.1	52.6
LS [7]	<u>90.0</u>	<u>76.8</u>	<u>32.6</u>	<u>53.5</u>
HML (Ours)	**91.3**	**78.6**	**33.6**	**54.1**

4 Experiments

4.1 Settings

Datasets. We conduct experiments on four datasets on multi-label classification, including VOC12 [14], COCO [28], CUB-200 (CUB) [41] and NUS-WIDE (NUS) [5]. VOC12 is a commonly used dataset on general objects. COCO contains a large number of small objects. CUB focuses on the fine-grained attribute identification of birds. NUS is a large-scale multi-label dataset. The details of datasets are listed in Table 1. For the single positive label setting, we use the same sampled labels from the original dataset as in [7].

Implementation details. We follow the same schedules in [7]. 20% data from the training set is collected for validation. We use ResNet-50 [16] as our backbones in all experiments. For each experiment, the model is trained for 10 epochs with Adam optimizer [22]. We do grid search on learning rates in $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and batch sizes in $\{8, 16\}$. The best model is selected by validation sets and used to compare with other methods. We integrated the proposed methods, including hyperspherical learning and adaptive learning, into our model, named hyperspherical multi-label classification (HML). The model is thoroughly evaluated in three scenarios (single positive labels, partial positive labels and full labels).

4.2 Single Positive Labels

In Table 2, we evaluate the proposed method in the single positive labels scenario. Our method comprehensively outperforms the methods proposed in [7].

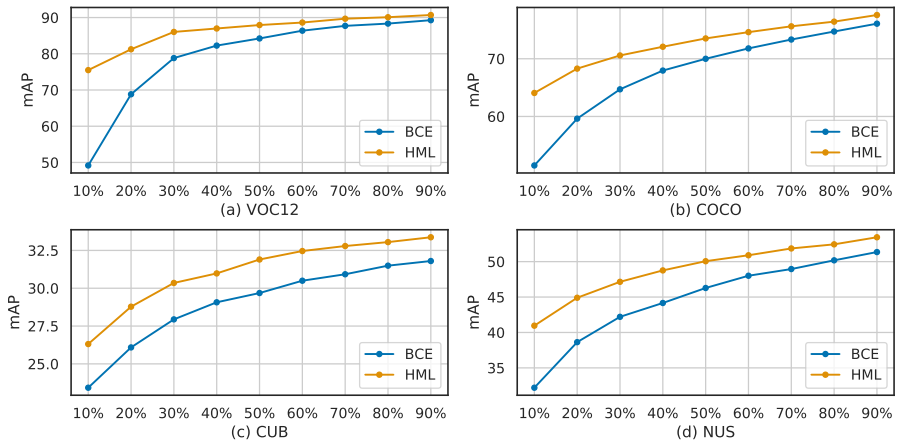

Fig. 4. Experiments on partial positive labels. The mAP is reported in different proportions of positive labels.

Especially for the COCO dataset, our method has an improvement of 3.8% compared to the WAN [7]. Previous methods do not consistently improve on all the datasets. By comparison, our method exhibits strong generalization on diverse datasets. These experiments verify the effectiveness of suppressing false negatives in single positive labels.

4.3 Partial Positive Labels

In order to further study the impact of false negatives, we conduct experiments on partial positive labels as shown in Fig. 4. In this scenario, impacts of different proportions of positive labels are evaluated. It is not surprising that the performance decreases significantly in very few positives. For general binary cross-entropy loss, the performance drops 40% in the VOC12 dataset with 10% positive labels. Compared to BCE, our method performs better with any proportion of positive labels. Also, the smaller the number of positives, the more significant the improvement. This result illustrates that our method can handle false negatives more effectively than the baseline method.

4.4 Full Labels

We also evaluate our method in full labels scenario in Table 3. Label smoothing (LS) [7] is a strong baseline in the full labels scenario compared to BCE, while our method surpasses LS by 1-2%. It shows that our method not only improves performance in datasets with single positive labels, but also generalizes well to datasets with full labels. In hyperspherical space, we could focus on learning positives and suppress the excessive contribution of false negatives.

Table 4. Ablation study on hyperspherical learning (HL) and adaptive learning (AL) in mAP metric.

		Single Positive Labels				Full Labels			
HL	AL	VOC12	COCO	CUB	NUS	VOC12	COCO	CUB	NUS
×	×	86.9	66.2	19.1	42.8	90.2	77.6	32.1	52.6
✓	×	**89.1**	70.4	20.6	46.5	**91.3**	78.4	33.5	53.6
✓	✓	**89.1**	**70.7**	**21.1**	**46.7**	**91.3**	**78.6**	**33.6**	**54.1**

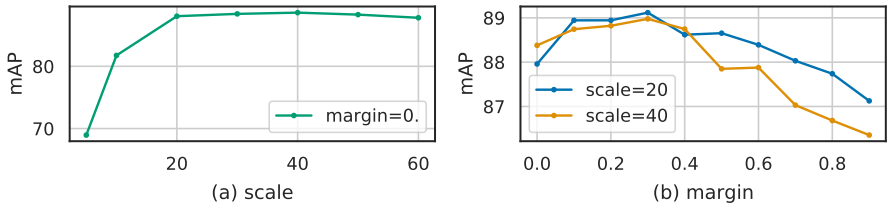

Fig. 5. Ablation study on scale and margin for VOC12 dataset in mAP metric.

4.5 Ablation Study

Learning in Hyperspherical Space. We study the impact of hyperspherical learning in scenarios of single positive labels and full labels in Table 4. In single positive labels scenario, our method significantly improves (up to 4.2%) over the baseline method. Even in the full-labeled dataset, our method has a consistent gain in performance (about 1%) across multiple datasets. It illustrates the strength of our approach in two ways. First, optimizing the cosine similarity facilitates multi-label learning. Second, hyperspherical learning is superior at suppressing the effects of noisy data.

Adaptive Learning. The adaptive variant of our method performs slightly better than the fixed variant as shown in Table 4. In the single positive labels scenario, the adaptive variant, in particular, outperforms 0.5 % in the CUB dataset. The CUB dataset is more challenging because more than 95% positive labels are discarded in this scenario. The adaptive variant effectively balances the impact of negatives and positives from the perspective of gradients, which makes the optimization easier.

Scale and Margin. The scale and margin are crucial hyperparameters in learning multi-label classification in hyperspherical space. An experimental study on ablation of the scale and margin is shown in Fig. 5. With no margin, the performance gradually converges with the increase of scale. It is because that small scale leads to large gradients even when cosine similarity equals 1. In order to comprehensively investigate the influence of margin, we conduct experiments on scale=20 and scale=40. With the increase of angular margin, the gain of performance is firstly strengthened and then weakened. The best performance is

Table 5. Ablation study on positives and negatives for VOC12 dataset in mAP metric.

Setting	mAP
No Margin	88.0
Only Positives	83.4
Only Negatives	88.2
Positives & Negatives	89.1

reached at mAP=89.1 with scale=20, margin=0.3. As analyzed in Fig. 3, the balance of gradients between positives and negatives is closely related to the choice of margin. The smaller margin could make the optimization of positives insufficient, while larger margin down weights contribution of negatives. It should be noted that this grid search of scale and margin is only conducted in single positive labels of the VOC12 dataset. The best hyperparameters above with scale=20 and margin=0.3 would also be used in other datasets' experiments for a fair comparison.

Margin on Positives and Negatives. The margin is only applied in positives in face verification during training. In our method, the margin is also added to negatives. To explore this difference, we experiment on the four settings of margin on positives and negatives shown in Table 5.

No Margin. It is the baseline with no margin on positives and negatives.

Only Positives. In this setting, the margin is only added on positives. Since annotations are unknown during testing, there is an inconsistency that the training uses the margin while the testing does not. In testing, the drop of margin reduces the probability of positives, making it difficult to distinguish positives from negatives. As a result, the *Only Positives* is 4.6% worse than *No Margin*. This inconsistency on positives leads to poor generalization.

Only Negatives. In this setting, the margin is only added to the negatives. On the one hand, the margin helps to reduce the impact of false negatives. On the other, the reduced probability would not degrade the ability to classify. The *Only Negatives* is slightly better than *No Margin*. It shows the importance of suppressing false negatives.

Positives & Negatives. In this setting, the margin is added on all samples and keeps the consistency of training and testing. The margin on positives helps to improve the cosine similarity between the feature and its corresponding positive label embedding, while the margin on negatives suppresses the impact of false negatives. *Positives & Negatives* exceeds *No Margin* 1.1% in mAP. It indicates the necessity to add the margin on both positives and negatives in optimizing multi-label classification in hyperspherical space.

4.6 Label Correlation

As discussed in Sec. 3.5, we estimate the label correlation in COCO dataset in Fig. 6. The higher is the correlation value, the stronger is the correlation

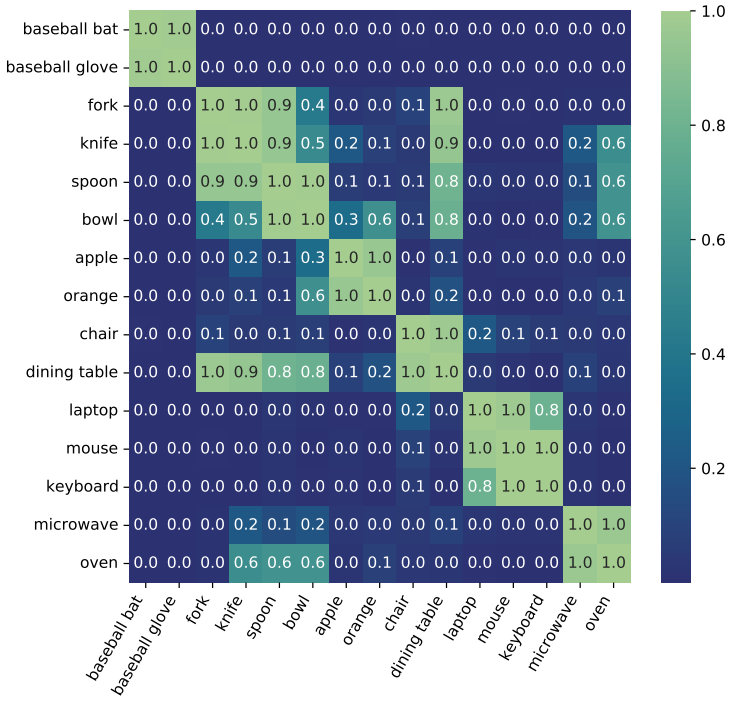

Fig. 6. Label correlation in COCO dataset. The labels with correlation value larger than 0.95 with other labels are shown.

between the corresponding two labels. We could see some labels sets with strong correlations: {baseball bat, baseball glove}, {fork, knife, spoon, bowl, dinning table}, {apple, orange}, {chair, dinning table}, {laptop, mouse, keyboard} and {microwave, oven}. All these combinations have a higher probability of appearing in the same scene. But the correlations between {apple} and {baseball bat, baseball glove} are weak. It is consistent with the rareness of their co-occurrence in the same scene.

5 Conclusion

In this paper, we present a novel perspective for learning multi-label classification in hyperspherical space. We thoroughly explore the impact of false positives in noisy multi-label tasks and propose the margin based sigmoid function for multi-label classification. To reduce manual hyperparameters searching, adaptive learning is incorporated into model optimization. Experiments show that our approach significantly improves performance in various scenarios, ranging from single and partial positive labels to full labels. In future work, we intend further to explore the problem of multi-label classification, as it is expected to be extended to the multi-label image retrieval task in hyperspherical space.

References

1. Akbarnejad, A.H., Baghshah, M.S.: An efficient semi-supervised multi-label classifier capable of handling missing labels. IEEE Transactions on Knowledge and Data Engineering **31**(2), 229–242 (2018)
2. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
3. Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R.: An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST) **12**(5), 1–32 (2021)
4. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019)
5. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009)
6. Cid-Sueiro, J.: Proper losses for learning from partial labels. In: Advances in neural information processing systems. pp. 1565–1573. Citeseer (2012)
7. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: CVPR. pp. 933–942 (2021)
8. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. The Journal of Machine Learning Research **12**, 1501–1536 (2011)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
10. Dong, H.C., Li, Y.F., Zhou, Z.H.: Learning from semi-supervised weak-label data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
12. Durand, T., Mehrasa, N., Mori, G.: Learning a deep convnet for multi-label classification with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 647–657 (2019)
13. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 213–220 (2008)
14. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
15. Fan, X., Jiang, W., Luo, H., Fei, M.: Spherereid: Deep hypersphere manifold embedding for person re-identification. Journal of Visual Communication and Image Representation **60**, 51–58 (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. He, X., Zemel, R.: Learning hybrid models for image annotation with partially labeled data. Advances in Neural Information Processing Systems **21**, 625–632 (2008)

18. Huang, S.J., Zhou, Z.H.: Multi-label learning by exploiting label correlations locally. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 26 (2012)
19. Huynh, D., Elhamifar, E.: Interactive multi-label cnn learning with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9423–9432 (2020)
20. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: NIPS. vol. 2, pp. 897–904. Citeseer (2002)
21. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1719–1726. IEEE (2006)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015), http://arxiv.org/abs/1412.6980
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
24. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16478–16488 (2021)
25. Li, Q., Peng, X., Qiao, Y., Peng, Q.: Learning label correlations for multi-label image recognition with graph networks. Pattern Recognition Letters **138**, 378–384 (2020)
26. Li, W., Wang, L., Li, W., Agustsson, E., Berent, J., Gupta, A., Sukthankar, R., Van Gool, L.: Webvision challenge: Visual learning and understanding with web data. arXiv preprint arXiv:1705.05640 (2017)
27. Li, X., Liu, B.: Learning to classify texts using positive and unlabeled data. In: IJCAI. vol. 3, pp. 587–592. Citeseer (2003)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
29. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
30. Liu, W., Zhang, Y.M., Li, X., Yu, Z., Dai, B., Zhao, T., Song, L.: Deep hyperspherical learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 3953–3963 (2017)
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (October 2021)
32. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
33. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia **22**(10), 2597–2609 (2019)
34. Meng, Q., Zhang, W.: Multi-label image classification with attention mechanism and graph convolutional networks. In: Proceedings of the ACM Multimedia Asia, pp. 1–6. ACM (2019)

35. Nguyen, N., Caruana, R.: Classification with partial labels. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 551–559 (2008)
36. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11557–11568 (2021)
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)
38. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)
39. Sun, Y.Y., Zhang, Y., Zhou, Z.H.: Multi-label learning with weak label. In: Twenty-fourth AAAI conference on artificial intelligence (2010)
40. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. journal (2011)
42. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
43. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1041–1049 (2017)
44. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
45. Wang, L., Liu, Y., Qin, C., Sun, G., Fu, Y.: Dual relation semi-supervised multi-label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6227–6234 (2020)
46. Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., Wen, S.: Multi-label classification with label graph superimposing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12265–12272 (2020)
47. Wen*, Y., Liu*, W., Weller, A., Raj, B., Singh, R.: Sphereface2: Binary classification is all you need for deep face recognition. In: 10th International Conference on Learning Representations (ICLR) (Apr 2022), `https://openreview.net/forum?id=l3SDgUh7qZO`, *equal contribution
48. Wojke, N., Bewley, A.: Deep cosine metric learning for person re-identification. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 748–756. IEEE (2018)
49. Xie, M.K., Huang, S.J.: Partial multi-label learning with noisy label identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
50. Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: European Conference on Computer Vision. pp. 649–665. Springer (2020)
51. Yu, F., Rawat, A.S., Menon, A., Kumar, S.: Federated learning with only positive labels. In: International Conference on Machine Learning. pp. 10946–10956. PMLR (2020)
52. Yu, Y., Pedrycz, W., Miao, D.: Multi-label classification by exploiting label correlations. Expert Systems with Applications **41**(6), 2989–3004 (2014)

53. Zhao, F., Huang, Y., Wang, L., Tan, T.: Deep semantic ranking based hashing for multi-label image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1556–1564 (2015)
54. Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., Li, J.: Transformer-based dual relation graph for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 163–172 (2021)
55. Zhao, J., Zhao, Y., Li, J.: M3tr: Multi-modal multi-label recognition with transformer. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 469–477 (2021)
56. Zhu, J., Liao, S., Lei, Z., Yi, D., Li, S.: Pedestrian attribute classification in surveillance: Database and evaluation. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 331–338 (2013)
57. Zhu, Y., Kwok, J.T., Zhou, Z.H.: Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering **30**(6), 1081–1094 (2017)